Stochastic Backpropagation

Presented By Sahil Manocha

Danilo J. Rezende, Shakir Mohamed, Daan Wierstra

2014

Exponential Families

$$p(x|\eta) = h(x)g(\eta)exp\{\eta^T u(x)\}$$

Maximum Likelihood Estimation

$$p(X|\eta) = \int h(x)g(\eta)exp\{\eta^T u(x)\}$$

Differentiating, we get

$$\frac{\partial p(X|\eta)}{\eta} = \nabla g(\eta) \int h(x) \exp\{\eta^T u(x) dx\}$$
$$+ g(\eta) \int h(x) \exp\{\eta^T u(x)\} u(x) dx$$

Exponential Families

Alternatively, $\eta = \eta(\theta)$,

$$p(x|\eta) = h(x) exp\{\eta^T u(x) - A(\eta)\}$$

A is the *log-partition function*.

Exponential Families

Theorem

The log-partition function $\theta \to A(\theta)$ is infinitely differentiable on its open domain $D := \{\theta \in \mathbb{R}^d : A(\theta) < \infty\}$. Moreover, A is convex.

Proof.

For convexity, let $\theta_\lambda = \lambda\theta_1 + (1-\lambda)\theta_2$, where $\theta_1, \theta_2 \in D$. Then, $\frac{1}{\lambda} \geq 1$ and $\frac{1}{1-\lambda} \geq 1$, and Holder's inequality is applicable. (Since the coefficients are conjugate exponents). We get,

$$log \int h(x) exp(\langle \theta_\lambda, u(x) \rangle) dx$$

$$= log \int h(x) exp(\langle \theta_1, u(x) \rangle)^\lambda exp(\langle \theta_2, u(x) \rangle)^{1-\lambda} dx$$

$$\leq \log \left(\int h(x) exp(\langle \theta_1, u(x) \rangle)^{\frac{\lambda}{\lambda}} dx \right)^\lambda \left(\int exp(\langle \theta_2, u(x) \rangle)^{\frac{1-\lambda}{1-\lambda}} dx \right)^{1-\lambda}$$

Exponential Distribution

Proof (Cont.)

$$= \lambda log \int h(x) exp(\langle \theta_1, u(x) \rangle) dx + (1 - \lambda) log \int h(x) exp(\langle \theta_2, u(x) \rangle) dx$$

$\square$

Exponential Distribution

Convexity makes estimation in exponential families substantially easier. Indeed, given a sample $X_1, \ldots X_n$ assume that we estimate θ by maximizing likelihood (equivalently, minimizing the log loss):

$$\min_{\theta} \sum_{i=1}^{n} \log \frac{1}{p_\theta(X_i)} = \sum_{i=1}^{n} [-\langle \theta, u(X_i) \rangle + A(\theta)]$$

which is convex in θ.

Stochastic Backpropagation

Gradient descent methods in latent variable models require computations

$$\nabla_\theta \mathbb{E}_{q_\theta}[f(x)]$$

$$\theta \sim q_\theta(.)$$

f = loss function

Quantity is difficult to compute:

1. expectation is unknown
2. indirect dependency on q

Bonnet's Theorem

Let $f(x)$: $\mathbb{R}^d \rightarrow \mathbb{R}$ be a integrable and twice differentiable function. The gradient of the expectation of $f(x)$ under a Gaussian distribution $\mathcal{N}(x|\mu, \Sigma)$ with respect to the mean μ can be expressed as the expectation of the gradient of $f(x)$.

$$\nabla_{\mu_i} \mathbb{E}_{\mathcal{N}(\mu,\Sigma)}[f(x)] = \mathbb{E}_{\mathcal{N}(\mu,\Sigma)}[\nabla_{x_i} f(x)]$$

Price's Theorem

Let $f(x)$: $\mathbb{R}^d \to \mathbb{R}$ be a integrable and twice differentiable function. The gradient of the expectation of $f(x)$ under a Gaussian distribution $\mathcal{N}(x|\mu, \Sigma)$ with respect to the covariance Σ can be expressed in terms of the expectation of the Hessian of $f(x)$ as:

$$\nabla_{\Sigma_{i,j}} \mathbb{E}_{\mathcal{N}(\mu,\Sigma)}[f(x)] = \frac{1}{2} \mathbb{E}_{\mathcal{N}(\mu,\Sigma)}[\nabla^2_{x_i,x_j} f(x)]$$

Backpropagation

By applying chain rule, we get:

$$\nabla_\theta \mathbb{E}_{\mathcal{N}(\mu,\Sigma)}[f(x)] = \mathbb{E}_{\mathcal{N}(\mu,\Sigma)}\left[\mathbf{g}^T \frac{\partial \mu}{\theta} + \frac{1}{2} Tr\left(H \frac{\partial \Sigma}{\partial \theta}\right)\right]$$

$\mathbf{g}$: gradient
H: hessian

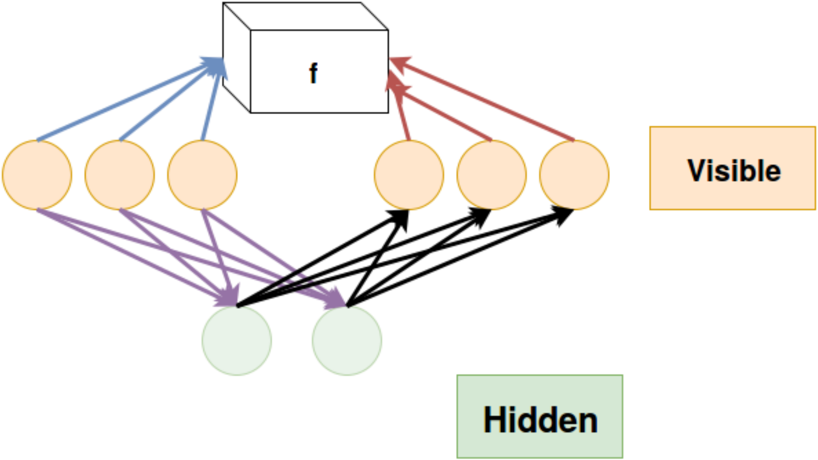
For General Distributions

$$\nabla_\theta \mathbb{E}_p[f(x)] = \mathbb{E}_{p(x|\theta)}[B(x)\nabla_x f(x)]$$

where B is a non-linear function.
For exponential distributions:

$$B(x) = \frac{[\nabla_\theta \eta u(x) - \nabla_\theta A]}{[\nabla_x log[h(x)] + \eta^T \nabla_x u(x)]}$$

Relation to RBM



GB-RBM

Energy Function

$$E(v,h) = \sum_{i=1}^{V} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \boldsymbol{c}^T \boldsymbol{h} - \sum_{j=1}^{V} \sum_{i=1}^{H} \frac{v_i}{\sigma_i} h_j w_{ij}$$

Probability

$$P(v) = \sum_{\boldsymbol{h}} \frac{1}{Z} e^{-E(\boldsymbol{v},\boldsymbol{h})} = \sum_{\boldsymbol{h}} \frac{1}{Z} e^{-\sum_{i=1}^{V} \frac{(v_i - b_i)^2}{2\sigma_i^2} + \boldsymbol{c}^T \boldsymbol{h} + \sum_{j=1}^{V} \sum_{i=1}^{H} \frac{v_i}{\sigma_i} h_j w_{ij}}$$

$$P(v) = \frac{1}{Z} e^{-F(v)}$$

where $F(v)$ is free energy

GB-RBM

Free Energy

$$F(v) = -log\Big(\sum_{\boldsymbol{h}} e^{-\sum_{i=1}^{V}\frac{(v_i-b_i)^2}{2\sigma_i^2}+\boldsymbol{c}_{\boldsymbol{h}}^T\boldsymbol{h}+\sum_{j=1}^{V}\sum_{i=1}^{H}\frac{v_i}{\sigma_i}h_j w_{ij}}\Big)$$

Simplifying the term within the *log*

$$\sum_{\boldsymbol{h}} e^{-\sum_{i=1}^{V}\frac{(v_i-b_i)^2}{2\sigma_i^2}+\boldsymbol{c}^T\boldsymbol{h}+\sum_{j=1}^{H}\sum_{i=1}^{V}\frac{v_i}{\sigma_i}h_j w_{ij}}$$

$$= e^{-\sum_{i=1}^{V}\frac{(v_i-b_i)^2}{2\sigma_i^2}} \times \prod_j \big(e^{c_j+\sum_{i=1}^{V}\frac{v_i}{\sigma_i}w_{ij}}+1\big)$$

Substituting, we get

$$F(v) = \sum_{i=1}^{V}\frac{(v_i-b_i)^2}{2\sigma_i^2} - \sum_j log\big(e^{c_i+\sum_{i=1}^{V}\frac{v_i}{\sigma_i}w_{ij}}+1\big)$$

GB-RBM

Conditional Probability

$$P(\boldsymbol{h}|\boldsymbol{v}) = \frac{\prod_j e^{\left(c_j h_j + \sum_{i=1}^{V} v_i w_{ij} h_j\right)}}{\prod_j \sum_{h_j} e^{\left(c_j h_j + \sum_{i=1}^{V} v_i w_{ij} h_j\right)}}$$

$$= \prod_j \frac{e^{\left(c_j h_j + \sum_{i=1}^{V} v_i w_{ij} h_j\right)}}{\sum_{h_j} e^{\left(c_j h_j + \sum_{i=1}^{V} v_i w_{ij} h_j\right)}}$$

$$= \prod_j P(h_j|\boldsymbol{v})$$

Conclusions

Stochastic Backpropagation possible as shown above.

References I

Danilo Jimenez Rezende, Shakir Mohamed, Daan Wierstra
Stochastic Backpropagation and Approximate Inference in Deep Generative Models